PostScript ....

Appropriate analysis and reporting of cluster randomised trials

Dyson *et al*[1] use a pragmatic design to address an interesting question, but I am concerned that the statistical analysis may be inappropriate and could have led to erroneous conclusions being drawn. The study is a cluster randomised controlled trial. Instead of randomising individual house officers (HOs), the authors have randomised groups of HOs (those working at the same hospital). This is entirely appropriate. As the authors point out, randomising individual HOs would risk contamination between the two study groups by HOs sharing aide memoires.

However, if groups, rather than individuals, are randomised then the use of standard statistical techniques may be inappropriate. These techniques assume that all observations (that is, all individuals) are independent of each other. Yet in a cluster trial this may not be true. HOs at the same hospital are likely to share characteristics and learning experiences, and thus be more similar to each other than HOs at different hospitals. Assuming independence in these circumstances may lead to an overestimate of statistical power of the study and an underestimate of the p value.

For this reason, cluster trials should be published with an estimate of the degree of clustering within groups (the intraclass correlation coefficient) and the effect that this has upon statistical power (the design effect). The potential effect of clustering should be considered in the sample size calculation and analysis should take potential clustering into account. The fewer groups randomised and the more individuals there are per group, the greater the potential impact of any clustering. This study entailed randomising eight hospitals, with presumably 15–20 HOs per hospital, so the potential effect of clustering should not be ignored.

Before we can accept the conclusions of this study we need some more information. What was the intraclass correlation coefficient for these data? How many HOs were included from each hospital? Was analysis undertaken at group (hospital) or individual (HO) level? If an individual level analysis was undertaken, was this adjusted for potential clustering?

Cluster trials are a valuable tool in emergency medicine research, and this study is a good example, yet care needs to be taken in statistical analysis and reporting. This issue has been addressed by the NHS Health Technology Assessment Programme,[2] the *BMJ*,[3] and the emergency medicine literature.[4] Guidelines have recently been published for reporting cluster trials,[5] we should ensure that articles in the *EMJ* follow them.

S Goodacre
Medical Care Research Unit, University of Sheffield, Sheffield, UK; s.goodacre@sheffield.ac.uk

References

1 **Dyson E**, Voisey S, Hughes S, *et al*. Educational psychology in medical learning: a randomised controlled trial of two aide memoires for the recall of causes of electromechanical dissociation. *Emerg Med J* 2004;**21**:457–60.
2 **Ukoumunne OC**, Guilliford MC, Chin S, *et al*. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technol Assess* 1999;**3**:issue 5.
3 **Campbell M**, Grimshaw J. Cluster randomised trials: time for improvement. *BMJ* 1998;**317**:1171–2.
4 **Wears RL**. Statistical methods for analyzing cluster and cluster-randomized data. *Acad Emerg Med* 2002;**9**:330–41.
5 **Campbell MK**, Elbourne DR, Altman DG, *et al*. CONSORT statement: extension to cluster randomised trials. *BMJ* 2004;**328**:702–8.

Authors' response

We must accept that our original analysis, which assumed statistical independence between observations obtained from staff within the same hospital, might not be justified. To explore this possibility we have computed intracluster correlation coefficients (ICCs) using estimated components of variance obtained from an analysis of variance in which hospitals were treated as random effects within a nested sampling design.

With regards the total score at 60 seconds the between hospital component of variance was negative and hence the estimated ICC was set to zero. The ICCs and variance inflation factors (VIFs, assuming an average cluster size of 15) for all four outcome measures are presented in table 1.

As pointed out, the consequences of positive ICCs is that the reported p values, which ignored the clustering effect, will tend to be biased downwards. A subsequent analysis, which adjusts for clustering within the study, produced increased p values for all outcomes with that for the score at 60 seconds remaining significant at the 5% level.

We did, however, state in the paper that the results were at best of marginal significance, statistically. The ceiling of a maximum of eight correct causes may have reduced the ability to show a significant effect, if one exists. Despite these p value discussions, the paper remains of importance for two reasons. Firstly, it points out that despite the best of intentions, the use of a device to augment recall may potentially lead to adverse effects; 78% house officers could recall hypothermia, which in the UK is an uncommon cause with a long treatment wheelbase, while only 35% remembered hypoxia, a more common cause with rapid treatment. Secondly, such devices may be subject to study of their effectiveness, albeit with difficulty.

P McQuillan
Department of Intensive Care Medicine, Queen Alexandra Hospital, Cosham, Portsmouth, UK

B Higgins
Division of Mathematics and Statistics, University of Portsmouth, UK

Correspondence to: Dr P McQuillan, peter.mcquillan1@ntlworld.com

Table 1 Intracluster correlation coefficients (ICCs) and variance inflation factors (VIFs) for all four outcome measures

Outcome	Original p values	ICC	VIF	Adjusted p values
Score at 60 seconds	0.034	0.000	1.00	0.048
Total score overall	0.067	0.011	1.15	0.095
Proportion with maximum score at 60 seconds	0.054	0.002	1.03	0.077
Proportion with maximum score overall	0.100	0.016	1.22	0.123